Does Where Words Come From Matter? Leveraging Self-supervised Models for Multilingual ASR and LID

Gaofei Shen
S4920155

August 17, 2022

Acknowledgement

Completing this thesis has been a journey. I don't think I would've made it without the support from so many people along the way. To **Ben**, you're the best emotional support Canadian on this planet. To **Amarens**, thank you for listening to my rants about Dutch burger places almost every time we hang out. To **Leslie**, thank you for all the cool ideas and positive energy. To **Sarah**, thank you for being the mid-western big sister I never knew I needed. I also owe the huge support from my parents. I would not have come this far without you two.

Many thanks to **Matt** and **Shekhar** for being on my thesis supervisory board and giving me input in editing my thesis. I am tremendously grateful for **Vass**'s support during my PhD application season. And of course, I thank all the professors and PhDs working on the Voice Technology program - **Xiyuan, Phat, Jelske**. I am also grateful to the entire Voice Tech program for supporting each other during the entire year. I wish all the best to every one of us so that we can chase our dreams.

Last but not least, I thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster.

Contents

Abstract

While end-to-end ASR systems have evolved to achieve great performance in monolingual speech recognition in many languages, researchers have tried to improve the performance of these systems further with several different approaches. For example, researchers have found potential ways to leverage the end-to-end architecture for multilingual code-switching speech recognition by fine-tuning pre-trained models on multilingual datasets directly (Lovenia et al., 2022). Because the previous attempts focused on higher-resourced language pairs such as Mandarin and English, this thesis tests if training end-to-end ASR systems based on self-supervised learning models with multilingual data directly can improve multilingual ASR performances for lower-resourced language pairs such as Frisian and Dutch as well. It was found that fine-tuning monolingual end-to-end models with code-switching dataset can achieve good results. Additionally, researchers have also found that the hidden representations generated by the intermediate layers in the neural network encode certain acoustic features (Pasad et al., 2021). This thesis also proposes using outputs from the intermediate layer to train a language identification system that can measure the language integration of code-switching utterances. Based on previous research on multilingual, code-switching capable ASR systems (Baevski et al., 2020; Bentum, 2022; Tseng et al., 2022; Yılmaz et al., 2016), a language identification system that can indicate the level of language integration of a word should be able to improve the accuracy of code-switching ASR further. However, as the experiment in this thesis revealed, a simple LID model for very similar language pairs such as Frisian and Dutch does not produce great results. It is possible that using a LID module in building a truly multilingual speech recognition software is not the best approach for languages that have many similarities. It also reveals more future topics in the multilingualism research field in finding out more features that human listeners use to identify the dialect or languages being spoken.

Chapter 1

Introduction

With the development of artificial intelligence and machine learning, the performance of Automatic Speech Recognition (ASR) systems has reached a high point for higher-resource languages. End users can easily rely on products such as Google Assistant, Alexa, or Siri to facilitate some of their daily tasks. While a lot of ASR frameworks brand themselves as "multilingual", what they mean is that the ASR systems can recognize languages from the list provided that it is the *only* language used in the utterance. However, the reality of daily language use for multilingual speakers is that these speakers frequently alternate the language they use when talking to a multilingual interlocutor. A *true* multilingual ASR system should be able to recognize multiple languages in the same utterance. To achieve true multilingual ASR, the system will need to be able to handle code-switching utterances.

Code-switching is the language use phenomenon where a language user alternates the use of different languages in the same utterance. As will be elaborated in the literate review, code-switching scholars have hypothesized that language switching is not a binary distinction between loanword (total assimilation) or code-switching (minimal assimilation) (Appel & Muysken, 1987). For languages that have been in language contact for long periods of time, it can be relatively challenging for even native bilingual speakers to determine if a *foreign word* is a loanword or a code-switching. Here the term *foreign word* is used very loosely in that it could include foreign vocabulary that entered the language at any point of the language development process. They may be a word that looks and sounds like they are from another language, i.e. retaining some features of the original language such as spelling or pronunciation; or they may be a word that entered the vocabulary for a long time that people don't even realize it has foreign roots. With the wide variety of foreign words present in any language, we should treat code-switching as a gradient phenomenon, a spectrum ranging from most integrated (loanwords) to least integrated (insertional code-switching). In the context of the present study, the language contact phenomenon contributes to the challenges in the language identification (LID) task for multilingual ASR systems: the ASR system has a hard time "picking" the optimal downstream language-specific ASR module to produce the correct transcription. The basic idea, essentially, is that if the LID module can identify the code-switching words correctly, it can help pick out the most suitable downstream ASR module so that the ASR system can produce more accurate transcriptions in code-switching speech.

Developing ASR systems with current machine learning methods is an extremely data-heavy task. There are a limited number of language pairs that can generate the data required for developing true multilingual ASR systems. This thesis will look at two language pairs: Mandarin-English and Frisian-Dutch. As two of the most resource-abundant languages in the world, Mandarin and English get a considerable amount of attention from universities and companies owing to the large bilingual speaker populations worldwide. Hence many of the new advancements in ASR and LID are first made in these two languages. On the other hand, language pairs with fewer speakers should not be ignored. The linguistic landscape in Europe is perfect for conducting research and development in this field. Frisian is a language spoken in the north of the Netherlands. While it holds the status of one of the Netherlands' official languages, it is undoubtedly an under-resourced language. Since the country speaks Dutch in most official settings and most Frisian speakers are only literate in Dutch, Frisian speakers need to code-switch to Dutch on the daily basis. As descendants of the West-Germanic language, Frisian and Dutch are not only closely related, but they also have gone through extensive language contact due to economical, political, and geographical factors. These factors all contribute to the significant challenges in measuring language integration.

While end-to-end ASR systems have evolved to achieve great performance in monolingual speech recognition in many languages, researchers have made several attempts in improving the accuracy of the ASR systems further. Language model decoding is a common method used in conjunction with the popular wav2vec 2.0 architecture (2020) to reduce the word error rates (WER). Specific to code-switching and multilingual ASR, researchers have also found potential ways to leverage the same end-to-end self-supervised learning architecture for multilingual code-switching speech recognition for higher-resource language pairs (Lovenia et al., 2022). This thesis tests if the same methodology, training end-to-end ASR systems based on self-supervised learning models with multilingual data directly, can improve multilingual ASR performances for lower-resourced languages as well.

Language identification (LID) techniques have been a popular approach to solving code-switching and multilingual speech recognition. Previous attempts in language identification models focused on manual feature engineering to generate speech representations suitable for the LID task. Recently, researchers have also found that the hidden representations generated by the intermediate layers in the neural network encode certain acoustic features (Pasad et al., 2021). This thesis also proposes using outputs from the intermediate layer to train a language identification system that can measure the language integration of code-switching utterances. Based on architectures of multilingual, code-switching capable ASR systems (Baevski et al., 2020; Bentum, 2022; Tseng et al., 2022; Yılmaz et al., 2016), a language identification module is expected to provide an indication on where a word lies on the language integration spectrum. If such an experiment is successful, the language identification module could be a valuable addition to multilingual speech recognition software in enhancing performance by decreasing the errors caused by words with similar pronunciation or loanwords.

This thesis is organized as follows. Chapter two provides an overview of the background literature. Chapter three proposes the method used in this thesis while also including the datasets used and replication works. Chapter four evaluate the results and discuss the implications. Finally, the thesis concludes with a brief summary of the work as well as future directions for code-switching speech recognition and language identification research.

Chapter 2

Literature Review

To structurally explain the background literature, this chapter is organized into four main sections. Section one provides a basic overview of the theoretical background for analyzing the code-switching phenomenon in linguistics, especially including research on the phonetics and phonology of code-switching. Additionally, the section also formally defines the technical terms used in this project such as *loanwords, code-switching, matrix language, embedded language, and language score.* Section two takes a deeper dive into the language pairs being investigated. Section three introduces the automatic speech recognition frameworks for traditional monolingual applications. Section four examines the issues in multilingual speech recognition while studying topics specific to the language pairs that this thesis focuses on. After a short summary of the chapter, the research questions and the the hypotheses are presented.

2.1 Code-switching

Code-switching (CS) is the language use phenomenon where a speaker alternates the use of languages in the same utterance. Due to the abundance of approaches in the CS field, we need to define some technical terms to avoid confusion. The *matrix language*, for the purpose of the present study, can be considered the language that comprises the majority of the utterance. It is sometimes also called *base language* in other literature (Muysken, 1995). The *embedded language* is the other language(s) in the utterance in addition to the matrix language. It can also be referred to as the *guest language.* For example, consider the sentence below:

(1) Sometimes I'll start a sentence in English 然后 用 汉语 说 完.
sometimes I'll start a sentence in English then use Mandarin say PERFECT
"Sometimes I'll start a sentence in English *and finish in Mandarin.*"

Inter-sentential and inter-clausal code-switches are relatively more straightforward in both their syntax and phonology. Example (1) is an inter-clausal code-switch adapted from the title sake sentence appeared in Poplack (1980). We can see that the Mandarin sentence occupies the second half of the sentence and sits on one of the sentence boundaries. Both inter-clausal and inter-sentential code-switching occupy a sentence boundary or an utterance boundary. Hence, it is possible for the speaker to follow English syntax and phonology in the first half and then switch to Mandarin syntax and phonology in the second half. Intra-sentential code-switching, however, is comparatively more complicated. CS researchers have proposed that code-switching is not strictly categorical. It may be more intuitive to consider the variety of code-switching to be on a spectrum.

(2) 你 会 主动 code-switch 吗?
you will active code-switch QUES?
"Do you actively *code-switch*?"

The italicized compound word is the embedded language, English, and the rest of the sentence is written in Chinese, the matrix language. Depending on the context, a common differentiating factor in distinguishing several kinds of code-switching is the *switch location*. Inter-sentential code-switching is the CS that happens between sentences, i.e. the speaker "switches" to another language after finishing a sentence. Inter-clausal code-switching is very similar to inter-sentential code-switching but the switch point is between clauses instead of complete sentences. Intra-sentential code-switching is a bit complex in that the switch could be just one word, the nonce-borrowing coined by Poplack et al. (1988), or it could be a phrase. Example (2) is a sentence with single-word code-switching. Because the code-switch segment is surrounded by matrix language, a complete switch to English syntax and phonology is not only not necessary but cumbersome for the speaker. But the speaker usually needs to mark the language switch for the listener to parse the multilingual sentence correctly.

In the past four decades, linguists have proposed several extensive syntax models for code-switching (Muysken, 1995; Myers-Scotton, 1993; Poplack, 1980). Table 2.1 compares the three main camps of code-switching syntax literature.

Myers-Scotton	Muysken	Poplack
ML + EL constituent	Insertion	(Nonce) borrowing
EL-islands		Constituent insertion
ML-shift	Alternation	Flagged switching
ML-turnover		Code-switching under equivalence
(Style shifting)	Congruent lexicalization	(Style shifting)

Table 2.1: Schematic comparison of code-switching and -mixing typologies in three traditions (Muysken, 2000)

Because this thesis primarily concerns the phonetic and phonology side of code-switching, we will not need to dive too deep into the syntax of code-switching. But it is nonetheless important that we establish a theoretical ground that different code-switched segments have different levels of integration into the matrix language. The Matrix Language-Frame (MLF) Model proposed by Myers-Scotton (1993) is especially suitable for describing intra-sentential code-switching with its granularity and consistency. The different levels of integration for each layer of the model can be considered to be a "spectrum". On the one side, a monolingual utterance sits at one end of the spectrum as total integration; on the other side, an utterance containing insertional code-switching, i.e. an inter-sentential code-switching utterance, sits at the other end of the spectrum as with no integration into the matrix language.

2.1.1 The Sounds of Code-switching

Code-switching scholars have long focused on the syntax and morphology side of the phenomenon. It was only not too long ago that the phonetics and phonology of code-switching were systematically studied. One of the fundamental works (Bullock, 2009) in the field presented the notion of phonology as a metric of lexical borrowing: established borrowings are usually highly integrated into the phonology of the matrix language. An example Bullock (2009) gave for such borrowing is the assimilation of "VapoRub" to *vivaporú* into Spanish. This borrowing would sit on the assimilated end of the **spectrum** of non-assimilated to assimilated forms. Bullock (2009) also pointed out that the main difference between CS and borrowing is that the source of the lexical items was different. Namely, CS segments come from the embedded language vocabulary whereas borrowing comes from the matrix language vocabulary.

Linguists also found that for bilingual human listeners, code-switching behavior seems to have a specific onset feature that enables the listeners to process the perception easier (Piccinini & Garellek, 2014; A. Shen et al., 2020). For example, A. Shen et al. (2020) found evidence suggesting that, in English-to-Mandarin code-switching, the fundamental frequency (f0) depends on the tone of the code-switched Mandarin segment and the location of the code-switched segment. Piccinini and Garellek (2014) showed that Spanish-English speakers produced different f0 contours for code-switched sentences and in the code-switching audio to anticipate a code-switch. While previous literature suggested a longer processing time for code-switch utterances, evidence from the experiment in Piccinini and Garellek (2014) indicates that listeners process Spanish-English code-switch utterances with prosodic cues as fast as monolingual English utterances. Similar observations of changes in vowel quality and stop Voice Onset Time (VOT) were made by other experimental studies (Elias et al., 2017; Muldner et al., 2019; Piccinini & Arvaniti, 2015).

If code-switch features such as prosody, vowel quality, and stop quality can help human listeners in their perceptions of multilingual speech, perhaps the same holds for machine speech recognition.

These topics in code-switching are very crucial for the development of true multilingual automatic speech recognition systems. ASR systems cannot be truly multilingual unless they can tackle all kinds of code-switching in addition to language-specific speech recognition. On a high level, there seems to be a parallel between the theoretical models of code-switching perception and multilingual speech recognition software. While earlier ASR software defaulted to monolingual operation, increasingly the successful ASR systems have multilingual capabilities built in. Starting from a crude combination of monolingual systems coupled with a language switch at the input to a multilingually trained hybrid ASR system to the current trend of joint LID and end-to-end architecture. The evolution of multilingual ASR mirrors the development in our understanding of the multilingual brain.

However, it must be noted that the current language identification approaches, even with streaming end-to-end speech recognitions, do not yet have the capability of leveraging the extra information from the previous few frames in the frame-level language identification. Perhaps one of the future directions the multilingual ASR community can take is to incorporate theoretically grounded code-switching perception research into account when designing the next generation of code-switching detection systems.

2.2 Speech Recognition for the Language Pairs Investigated

2.2.1 Mandarin & English

Mandarin and English are two of the most resource-abundant languages in the world. Both languages are spoken widely in the world and both boast extremely high numbers of native (L1) speakers and second language (L2) speakers: 919,856,040 L1 speakers and 198,728,000 L2 speakers for Mandarin; 372,862,090 L1 speakers and 1079,609,320 L2 speakers for English (Eberhard et al., 2022). In the meantime, with a large number of Mandarin-speaking students and workers living in English-speaking countries, we can assume that there are more L1 Mandarin L2 English (L1M-L2E) speakers than L1 English L2 Mandarin (L1E-L2M) speakers.

Mandarin and English belong to different language families, Sino-Tibetan and Indo-European respectively, and went through drastically different historical development. There are a couple of crucial differences between the two languages. While Mandarin is a tonal language and has four tones, English is not. The two languages also have drastically different writing systems. Both Mandarin and English also went through extensive language contact with other languages in the world and hence have a large borrowed vocabulary in both languages.

A large overlap between Mandarin and English speakers facilitates code-switching between the two languages. As pointed out by H. Liu (2019), code-switching between Mandarin and English in China is connected to the large population of L2 English learners in China and the cultural contact between Chinese people and the Western world. While there is no doubt code-switching is a language behavior that is impossible to avoid, the attitude toward Mandarin-English code-switching is not always positive. For example, some Mandarin speakers may think that Mandarin-English code-switching can harm language integrity. Different population groups may also hold different attitudes toward CS in their daily lives due to exposure, proficiency in English, etc. These attitudes can be found in speakers of any language in the world. However, these studies on code-switching attitudes focused on code-switching from Mandarin to English, the other direction, from English to Mandarin was less studied. Anecdotal evidence suggests that while code-switch from Mandarin to English is common both in and outside of China, code-switching from English to Mandarin sounds less natural to proficient speakers. And code-switching from English to Mandarin is limited to certain specific scenarios such as language instruction settings or cultural exchange settings (i.e. talking about Chinese food). This discrepancy in attitude could be attributed to the fact that there are significantly more L1M-L2E speakers in the world.

2.2.2 Frisian & Dutch

Frisian and Dutch are two languages spoken in the Netherlands. The speaker population of Frisian and Dutch is much smaller in sharp contrast to Mandarin and English: 873,000 Frisian speakers and 16,000,000 Dutch speakers (Eberhard et al., 2022). Language resources such as corpora for the two languages are also more sparse than the more resourceful pair. But with the intertwined history, Frisian-Dutch code-switching is very much worth investigating.

Frisian was spoken widely along the coast of the North Sea. As the figure shows below, Frisian is regarded to be the closest relative to English in the West Germanic languages whereas Dutch (Netherlandic) and German are grouped into a sister cluster called Netherlandic-Germanic. However, it is important to keep in mind that languages and dialects do not have clear borders. When you cross the border from the Netherlands to Germany, for example, the language spoken by the locals does not make the big jump from Standard Dutch to Standard German. While its historical significance is notable, at present, only a small number of speakers

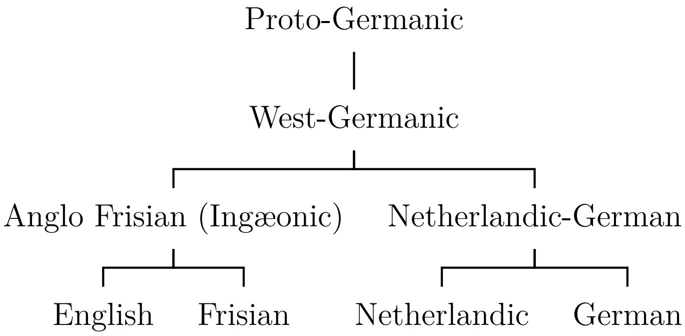

Figure 2.1: The traditional language family tree with Frisian, Dutch, German, and English

still remain. There are three main dialects of Frisian: West Frisian, spoken in the Friesland province in the Netherlands, East Frisian (Saterland Frisian), and North Frisian, spoken in small parts of North Germany. We must make the distinction that **Frisian** in this thesis from here on strictly refers to West Frisian, as the other two Frisian varieties are not mutually-intelligible with West Frisian. Hence the other two dialects are not included in the dataset or supported by the ASR software.

Descended from West-Germanic, Frisian bears similarities with English, Dutch, and German. While English and Frisian branched off as Anglo-Frisian from Dutch and German (Netherlandic-German), due to the geographical proximity to Dutch, Frisian received a significant amount of loanwords in various aspects of the language throughout its development. Gooskens and Heeringa (2012) proposed using Levenshtein distance to measure the distance between various dialects of Frisian with other Germanic languages. Gooskens and Heeringa (2012) found that, while Frisian shares a genetic relationship with Old English historically, it has grown closer to Dutch more than any other language in the West-Germanic group. The Town Frisian variant spoken in Leeuwarden, the capital of Friesland, is even more similar to Dutch than other Frisian dialects. This finding shows that language contact plays a more important role in languages being similar to each other than the genetic relationship in historical linguistics.

The interwoven relationship between Frisian and Dutch means code-switching between the two languages is not just common but also necessary in some situations. With conservation efforts by the Friesland provincial government and the Dutch national government, Frisian is no longer on the verge of extinction. However, for a long time, Frisian was considered a *lingua rustica* that carries less prestige compared to Dutch (Markey, 1980). Despite its relatively small speaker population, however, Frisian still has a considerable amount of dialects, possibly due to a lack of standardized written form for a long period. Most of the speakers can only speak

Frisian and only a small percentage can write in Frisian. Frisian gained its official language status in Friesland in 1956. It is possible to use Frisian in official settings and schools from then. This helped the preservation of the language. There are Frisian-instruction schools in the province so that young children can learn Frisian in an immersive setting. There has been an effort on producing Frisian textbooks, however since there have already been Dutch standardized textbooks, the textbook used by schools that engage in Frisian instructions are still in Dutch ("Language Plan Frisian 2030 –Frisian, for Now and Later," n.d.).

2.3 Relevant Automatic Speech Recognition Technologies For The Present Study

An informal extrapolation from the state of the art reveals the following few ASR technologies. The following ASR toolkits and frameworks also serve as the technical backbone of the methodology used in this thesis.

The Hidden Markov Model Toolkit (HTK), by its namesake, is a toolkit for researching and developing hidden Markov models (HMM). Hidden Markov models are probabilistic models that can efficiently use statistics to reduce the model size and improve speed. While HMMs have been around for decades, HTK still forms the backbone of popular tools used in phonetics research such as the FAVE forced alignment tool (Rosenfelder et al., 2015). First introduced in 2009, Kaldi is a research-focused toolkit for speech recognition (Povey et al., 2011). Based on Gaussian Mixture Models (GMM), Kaldi GMM and HMM are frequently used together as GMM can be used to generate the output probability of the HMM process. The recipe-based development system in Kaldi also helped in the quick iteration and replicating methods for different datasets. In addition to the traditional GMM models, Kaldi recipes can also make use of deep neural network (DNN) models to achieve even better performance.

PyTorch (Yang et al., 2022) is a popular machine learning library in Python. In addition to this lower-level library, there are more user-friendly higher-level machine learning libraries relying on the PyTorch implementation such as fastai and Transformers from HuggingFace. All of these Python libraries enabled researchers to quickly and easily iterate and experiment on newer ASR frameworks. These machine learning libraries have integration with the popular wav2vec 2.0 model (Baevski et al., 2020) for ASR and audio research. HuggingFace, apart from providing higher-level machine learning libraries suitable for speech recognition research, also hosts open-access datasets and pre-trained models. The transformers library (Wolf et al., 2020) builds on top of the frameworks such as PyTorch and Tensor-

Flow to enable quick evaluation, testing, and iteration for machine learning models. The standardized dataset format and streamlined inference tools also make ASR development considerably more accessible. This thesis not only uses the Common Voice dataset, the ASCEND dataset (Lovenia et al., 2022) and multiple pre-trained models published on HuggingFace, but hosts the completed ASR demonstrators on HuggingFace as well.

End-to-end models and self-supervised learning are two of the most recent trends in many subfields of machine learning. The wav2vec 2.0 model (Baevski et al., 2020) is an end-to-end model that was pre-trained using unannotated speech data for downstream tasks. The basic architecture of the model consists of a multi-layer convolutional network to extract features and then feeds the feature representations through a transformer architecture before passing the intermediate outputs through a quantization module (Baevski et al., 2020). Baevski et al. (2020) compared the performance of the model architecture with 12 transformer blocks (the base model) and 24 transformer blocks (the large model). Both models achieved great performances but the base model consumed significantly less time and computational resources in the pre-training process. Using un-labeled data such as the Librispeech corpus to pre-train the model first, the wav2vec 2.0 model demonstrated that self-supervised pre-training could provide great performance improvements in downstream ASR tasks while also reducing the amount of labeled data needed. These advantages make wav2vec 2.0 a great model for implementing ASR in low-resource languages. The Cross-lingual Speech Representation (XLSR) model is a pre-trained model with the wav2vec 2.0 Large architecture with a 53-language multilingual dataset combined from the Common Voice, BABEL, and Multilingual LibriSpeech corpora. Many wav2vec 2.0 models fine-tuned for higher- and lower-resourced languages alike used the XLSR model as their fine-tuning base. These fine-tuned models achieved great performance without needing a large amount of training data thanks to the self-supervised pre-training. This thesis will leverage the XLSR model and its fine-tuned derivatives as well.

Regarding the future development in ASR technologies, Hannun (2021) made the observation that semi-supervised learning such as the techniques used in wav2vec models (Baevski et al., 2020) is likely to improve on the state-of-the-art. Self and semi-supervised learning can leverage a large amount of un-annotated data, somewhat avoiding the data scarcity issue faced by many subfields of machine learning. Hannun (2021) also pointed out that the current performance metric for ASR systems, Word Error Rate (WER), and its various derivatives, will not be as impactful as it was before. While the last point may hold true for higher-resourced languages, I personally think the metric still has use in evaluating and testing ASR systems for lower-resourced languages.

2.4 Multilingual ASR

Traditionally, Automatic Speech Recognition systems can recognize utterances in one language. With the advancements in the field, multilingual Automatic Speech Recognition (MultiASR) is becoming a reality with several different implementations. The earliest attempts at MultiASR implemented a language identification module serving as a switch to feed the input data to a combination of several monolingual ASR systems. When correctly implemented, this kind of system can yield great results by virtue of the great monolingual performances of the individual ASR systems. But the system increases in size linearly with the number of languages added. As end-to-end ASR gained popularity, newer approaches started investigating single models trained on multilingual datasets.

We have seen newer frameworks such as Deep Speech (Hannun et al., 2014) and Deep Speech 2 (Amodei et al., 2015) that claimed great performances in more than one language under monolingual testing scenarios. Although the architecture is slightly out-of-date right now, we can still say that early end-to-end models such as Deep Speech 2 can perform great for monolingual utterances in the languages it was trained on (Mandarin and English). However, due to the unbalanced nature of training datasets, the result is often not as good as monolingual ASR systems.

Most recently, Tseng et al. (2022) proposed using self-supervised learning (SSL) pre-trained models such as wav2vec 2.0 (Baevski et al., 2020) to serve as speech representation extractor in addition to the classic Connectionist Temporal Classification (CTC) end-to-end speech recognition framework.

In addition to using the final output from a pre-trained model as input for language identification modules, there might be another more efficient way of obtaining language features from the raw audio signal. The intermediate transformer layers of the wav2vec 2.0 model were used as encoders in automatic speech recognition tasks. While literature studying the intermediate representations of the wav2vec 2.0 model is sparse, Pasad et al. (2021) found that the representations generated by each wav2vec 2.0 intermediate layer each encode slightly different information. The information encoded follows a trend from low-level, such as phonetic information, to high-level, word meaning. For example, Fan et al. (2021) investigated speaker verification and language identification using the representations generated from these intermediate layers as input features. Fan et al. (2021) found that lower layers can better distinguish speakers or languages.

2.4.1 Battling data scarcity

Code-switching detection has always been a hard topic in ASR. Several big challenges come with the topic. One of the biggest challenges is the lack of data. It is very hard for people to produce natural monolingual sentences in normal circumstances already, as much sociolinguistic research has shown. It is even harder to elicit code-switching bilingual speech. Such difficulty is magnified in lower-resourced languages. To solve the data scarcity issue, researchers adopted several different ways to increase the effective data size such as data augmentation (Yilmaz et al., 2018). In addition to the classic data manipulation techniques, Chang et al. (2019) proposed using Generative Adversarial Networks (GAN) with Reinforcement Learning (RL) to generate realistic code-switching sentences for data augmentation. While Chang et al. (2019) focused on generating new code-switching text data, it would be possible to generate code-switching audio data to facilitate the development of code-switching speech recognition systems. With limited resources and time, this thesis will not attempt using a similar method to generate more code-switching audio but leave the topic for future researchers to explore. However, as mentioned before, self-supervised learning models such as wav2vec 2.0 can leverage unlabeled data and cross-lingual training to make the best use of the limited labeled datasets. Hence this thesis will focus on the various applications of the wav2vec model.

2.4.2 Mandarin-English Speech Recognition

Software companies based in countries where Mandarin and English are spoken contributed to some of the most influential driving forces in the development of ASR and TTS software. Researchers from companies like Baidu, Google, Meta, Apple, and academic institutions have been working on datasets and algorithms since the beginning of the field. For example, the DeepSpeech 2 model from 2015 (Amodei et al., 2015) obtained similar or better performance compared to human transcriptionists in many speech-to-text tasks including read speech, accented speech, noisy speech in English and Mandarin. Granted, the authors only tested the model with monolingual utterances from different corpora, we can see the great potential in developing a true multilingual speech recognition system for Mandarin-English code-switching speech. However, with the trend to use more and more data in neural network training, we have to note that code-switching dataset is still a niche in the field. Researchers focusing on Mandarin-English bilingual code-switching speech recognition and language identification algorithms have relied on a limited number of datasets such as the SEAME (D.-C. Lyu et al., 2010).

As mentioned in the sections before, Mandarin and English have seen more resources in the development of multilingual speech recognition implementations. However, as resource abundant as the Mandarin and English languages are, most of the earlier ASR frameworks focused on getting better performance in monolingual speech recognition tasks. After all, it was not until relatively recently (the past decade (Amodei et al., 2015; Hannun et al., 2014)), that the monolingual performance of ASR systems become comparable to expert human transcriptions. The SEAME corpus (D.-C. Lyu et al., 2010) was one of the first code-switching corpora that were used for ASR development. While there are other Mandarin-English code-switching corpora such as H.-P. Shen et al. (2011), the availability of corpora was not guaranteed. The next chapter will dive deeper into the dataset availability issue further.

2.4.3 Frisian-Dutch Speech Recognition

In contrast to the Mandarin-English pair, Frisian-Dutch receives less attention from companies. Most of the developments of Frisian-Dutch ASR are led by non-profits, the regional government, and research institutions. However, the development of the ASR field in the direction of low-resourced languages brought several exciting advancements in Frisian-Dutch training datasets and speech recognition systems. The Fryske Akademy, the scientific research center focusing on Frisian culture, has undertaken multiple projects on speech technology applications for Frisian. There is a project on Dutch-Frisian code-switching speech recognition on the Frisian Audio Mining Enterprise project (FAME!) (Yilmaz et al., 2018). The corpus consists of radio broadcasts in Dutch and Frisian. More recent attempts in improving ASR performance for the language pair by (Bentum, 2022) created code-switching language tags for the Frisian Council Meeting Corpus (FCMC) to train a bilingual speech recognizer. In addition to the two academic-focused Frisian-Dutch corpora, Mozilla's Common Voice project also has both Frisian and Dutch recordings publicly available for experiments.

The FAME! project (Yilmaz et al., 2018) collected speech data of the Omrop Fryslân. A speech recognition system for transcribing the speech in radio broadcasts was developed. The system is capable of recognizing the bilingual code-switching speech of the local broadcast anchors. In 2021, the Frisian Council Meeting Corpus (FCMC) project (Bentum, 2022), under the Fryske Akademy's consultant, was created for the speech recognition system for its namesake. Along with more data, the new corpus also tackled challenges such as speech recognition in domain-specific settings (i.e. government debates) and extensive code-switching. The Common Voice corpus, on the other hand, only includes monolingual utterances while benefiting from crowd-sourcing data from enthusiastic native speakers of Frisian and Dutch. The methodology chapter will provide more complete descriptions of the corpora used.

The most recent implementations of Frisian-Dutch multilingual speech recognition systems by Bentum (2022) and Yılmaz et al. (2019) used Kaldi (Povey et al.,

2011) to quickly iterate their development of the ASR software. In the results presented by Bentum (2022), the authors mentioned that some of the Word-Error-Rate (WER) metrics are slightly misleading in the context. They argue that while the best WER achieved in the experiment is still relatively high at 29.59%, some portions of the errors are so minor that they should be excluded from the metric. Three categories of these minor errors were presented in the study: Compound words, Spelling variations, and Abbreviations. To better understand why the authors consider these errors minor, we will examine each one in detail below:

Since the FCMC corpus consists of government meeting recordings, a lot of domain-specific compound words in the legal field appear in speech. The ASR system sometimes breaks compound words up into their components. While this behavior does not hinder the comprehensibility of the transcript produced by the ASR system, the performance evaluation metric takes it into account. Secondly, due to the similar pronunciation of certain Frisian and Dutch words, the ASR system could attribute an erroneous language tag to the word and is penalized on performance metrics. The absence of a uniform spelling system also contributes to spelling variation errors. As mentioned in Markey (1980), Frisian, even to this day, still has quite a few inter-dialectal variations. With variations in spelling leaning toward a more “Frisian” spelling versus a more “dutch” spelling, Bentum (2022) argue that the language model cannot possibly take all the variations of the same word into account. Lastly, abbreviations are a classic challenge for ASR systems to conquer due to the variations in the pronunciations. The error metric could simply be a slight difference between the automatic transcription and the human transcriptions.

2.4.4 Language Identification Task

At the moment, end-to-end speech recognition frameworks, even for resourceful languages such as Mandarin and English, do not have great performance for code-switching speech compared to monolingual ASR as shown in Lovenia et al. (2022). Developing language identification methods to improve code-switching speech recognition could prove to be a fruitful attempt for researchers in the field. In the past decade, the trend is to use a mix of acoustic features and language models to improve the accuracy of language identification tasks and hence improve code-switching speech recognition performances. Attempting to solve the issue of code-switching detection in Mandarin-English speech recognition, Zhang (2012) first proposed to integrate the LID module into an already trained Gaussian Mixture Model (GMM) speech recognizer. Zhang (2012) used features extracted from the wave audio files to get a language feature GMM model and trained a support vector machine (SVM) as a LID module. Yilmaz et al. (2018) used data augmentation techniques to increase the code-switching data available for training a code-switching ASR system. Yılmaz et al. (2019) leveraged multi-graph decoding techniques to decrease the reliance on the language models and hence get better performance on the monolingual utterances.

More recently, with the shift to end-to-end ASR systems, newer language identification techniques emerged. Li et al. (2019) joined the popular CTC ASR model with a LID system and confirmed such a method could benefit code-switching ASR performance. Stacking 80-dimensional Mel filterbank energies on three frames at one time to train the CTC model and the LID module, the model trained on code-switching speech data was able to achieve 84% switching detection accuracy. Basing on the findings of Li et al. (2019), Tseng et al. (2022) first proposed leveraging self-supervised speech representation models and CTC end-to-end ASR frameworks to improve ASR performance for code-switching tasks. Tseng et al. (2022) found that self-supervised learning models such as wav2vec 2.0 can learn hidden representations that contain language identities. And this joint implementation can improve Mandarin-English code-switching speech recognition. This method has a few advantages over the previous attempts: 1. It can conduct frame-wise LID; 2. It showed that a model pre-trained on other languages can also encode language information for drastically different languages; 3. Additional language identification modules can still be beneficial to end-to-end ASR models. To conduct frame-level LID, Tseng et al. (2022) used Montreal Forced Aligner (McAuliffe et al., 2017) to find the word boundary and jointly trained the LID module with the CTC module to test the code-switching ASR performance. Tseng et al. (2022) compared using a simple fully connected (FC) layer and a bidirectional Long Short-Term Memory (BLSTM) as the language ID prediction head. BLSTM performed significantly better (20%) than the simpler structure of the FC layer.

Implementing a LID module similar to Tseng et al. (2022) in Frisian-Dutch would help us see if SSL speech representations can distinguish between very similar languages. Following similar implementations of MultiASR in more resourceful language pairs, I believe that the ASR performance for code-switching bilingual Frisian-Dutch speech could also be improved.

Summary

This chapter first defined the technical terms used in this thesis. I also provided an overview of the code-switching literature, especially focusing on the sounds of code-switching. The second section briefly introduced the language pairs investigated in this thesis. Section three outlined some of the relevant ASR implementation frameworks. Section four touches on one of the biggest challenges in multilingual

ASR and mentions previous ASR research on Mandarin-English and Frisian-Dutch as well as presenting the audio-based language identification methods. This chapter brings us on the path to the methods followed in this thesis.

2.5 Research question

Following the literature background explained earlier in the chapter, the research question for the thesis is two-fold:

1. Does fine-tuning an end-to-end ASR system with a multilingual dataset bring performance benefits to under-resourced language pairs such as Frisian and Dutch?

2. Can intermediate layers of the end-to-end neural network architecture in ASR be used to train a language identification system for closely related languages such as Frisian and Dutch?

Hypotheses

1. Following Lovenia et al. (2022), I hypothesize that the performance of a Frisian-Dutch bilingual code-switching capable speech recognition system based on wav2vec 2.0 framework (Baevski et al., 2020) could benefit from direct training with a multilingual corpus such as the FAME! corpus (Yılmaz et al., 2016). If the experiment result suggests otherwise, it is possible that the significantly more data afforded by the resourceful languages provided the performance enhancement.

2. Following Bullock (2009), Piccinini and Garellek (2014), A. Shen et al. (2020), and Tseng et al. (2022), and Pasad et al. (2021), because there are human recognizable patterns in code-switching between languages, sometimes native speakers have an intuition of whether a word is a loanword or code-switching. However, for language pairs that are as closely related as Frisian and Dutch, such phonetic cues may not be enough for the native speakers to make the correct judgment. I hypothesize that novel machine learning methods may be able to help us identify the features of code-switching transition. Intermediate representations from a deep neural network can be used to train a language classification model. If the results suggest otherwise, it could be due to the features generated by the network are not suitable for the language identification task. It could also suggest that native speakers rely on more than just acoustic features to perceive loanwords and code-switching.

Chapter 3

Datasets and Methodology

This chapter examines the datasets and outlines the methodology used in this thesis. the first section describes the datasets used in this thesis. Section two introduces the methods for investigating multilingual speech recognition. Based on previous research on Dutch-Frisian code-switching and Mandarin-English code-switching speech recognition, I adopted modern ASR implementation techniques and frameworks such as Deep Neural Network (DNN), and wav2vec 2.0. I followed the Kaldi (Povey et al., 2011) recipe used in the FAME! project (Yılmaz et al., 2016, 2019) and the FCMC project(Bentum, 2022). I also followed a similar training methodology used by Lovenia et al. (2022) to improve bilingual speech recognition performance by fine-tuning the wav2vec 2.0 model (Baevski et al., 2020) directly with bilingual speech. The last section explains how the intermediate representations from the previously mentioned wav2vec 2.0 models are used to train a language identification system.

3.1 Data

Unlike monolingual speech corpora, there is a significantly lower number of multilingual and code-switching corpora available for speech technology research and product development. On the Mandarin-English front, the SEAME corpus (D.-C. Lyu et al., 2010) has been the corpus of choice for Mandarin-English code-switching researchers for over a decade. This is because it was the only consistently available Mandarin-English code-switching corpus for a long time. Lovenia et al. (2022) pointed out that, while there were more Mandarin-English code-switching corpora such as CECOS (H.-P. Shen et al., 2011) in the past, many were no-longer publicly available due to communication chains for corpus access being broken, or the corpora were only developed for specific purposes. On the bright side, the Hong Kong University of Science and Technology released a new corpus called A Spontaneous Chinese-English Dataset (ASCEND) (Lovenia et al., 2022). ASCEND comprises

10.62 hours of spontaneous speech recorded in a casual environment in a speech lab. The corpus contains both inter-sentential and intra-sentential code-switching. Of all ten hours of recorded speech, they found roughly half contained code-switching.

For the lower-resourced pair, the situation is slightly different. While the resource is limited, the corpus development process is more organized thanks to the Fryske Akademy. The first Frisian-Dutch corpus, FAME! (Yılmaz et al., 2016) was released by researchers affiliated with the Fryske Akademy in 2016. The corpus consists of around 11 hours of speech, 8 hours of which are Frisian and the rest in Dutch. The annotation of the corpus, apart from transcriptions, also included speaker labels. Yılmaz et al. (2016) noted that there are a total of 3837 code-switching utterances in the entire speech corpus including inter-sentential and intra-sentential CS. The majority (75.6%) of these CS utterances are Frisian speakers switching to Dutch in a Frisian majority utterance. A very small percentage (2.5%) is CS in the other direction. The remaining bunch is what Yılmaz et al. (2016) call a *mixed-word.* Yılmaz et al. (2016) defined the term as neither Frisian nor Dutch.

The Frisian Council Meeting Corpus (FCMC) (Bentum, 2022), compiled in 2020, fills the vacuum of spontaneous speech dataset by collecting audio samples from government council meetings in Friesland. It includes 26 hours of Frisian speech and 23 hours of Dutch speech totaling 49 hours of recordings. The company Humain'r provided the transcription for the development of a multilingual ASR system to use in the council meeting environment. Due to the nature of spontaneous speech and the recording environment, the audio data is significantly noisier compared to lab recordings. The manual transcriptions provided by the FCMC corpus contain 44 data cleaning tags, which we used to clean up the transcription text. Removing unrecognizable audio is crucial for the training of a successful ASR system. Doing so removed one-third of the dataset, reducing the total size from 49 hours to 33 hours of speech. The transcriptions, instead of including cleaning labels and speaker labels like the FAME! corpus, contain language labels, borrowing labels, as well as code-switching direction labels identifying the matrix language and the embedded language. To provide these additional labels, Bentum (2022) developed a text-based language identification system trained on the FCMC transcriptions that 'tags' each word that was recognized. They also combined the language model trained on FAME! and FCMC to form an interpolated LM for better performance without retraining the model on a combined dataset.

3.2 Replication

I want to replicate findings in previous literature and establish them as a baseline for comparison with the original experiment. This section will roughly outline the method used in the previous implementations of Frisian-Dutch ASR and Mandarin-English ASR and how and if I was able to replicate their results.

3.2.1 Frisian-Dutch ASR with Kaldi

Following the Kaldi (Povey et al., 2011) recipe provided by the FAME! and FCMC projects, I replicated the performance they found in the Frisian-Dutch multilingual ASR system. Yılmaz et al. (2016) used 39-dimensional MFCC features including the delta and delta-delta features to train a DNN network in Kaldi. Bentum (2022) adopted a similar approach while improving the performance of the original DNN-based ASR system by using a more recent tDNN training. While most of the codebase was present along with the dataset, or publicly available, replicating both implementations have proved challenging. Neither project included sufficient documentation that enables easy replication. I was not able to re-trace their work step-by-step.

With regards to the wav2vec 2.0 models, I found a publicly available pre-trained wav2vec 2.0 XLSR fine-tuned on Frisian by Wietse de Vries. Initially, I wanted to fine-tune the model from scratch again with the updated Common Voice datasets since the dataset has received more data since a year ago. However, the public model still performed better than the model I fine-tuned. During the training process, I found that the model I trained performed significantly worse with a WER of over 30% compared to the quoted WER of 16.25% of the public model. Since no documentations exist on how the public Frisian model was fine-tuned, it seems unlikely that I can replicate or exceed the existing performance in the short timeframe of this thesis. Hence I opted to fine-tune the public model directly with FAME! data. In addition to just using the CTC module in wav2vec 2.0 in the final decoding step, I also wanted to test the performance of using a language model in wav2vec 2.0 model decoding. While the FAME! corpus included a language model trained for the Kaldi implementation, due to the discrepancy in language model formats, I decided to create a new n-gram language model specifically on the text data available to me.

3.2.2 Mandarin-English ASR with wav2vec 2.0

As mentioned in the literature review chapter, Lovenia et al. (2022) offered an interesting direction in developing multilingual ASR systems: fine-tuning wav2vec 2.0 framework with bilingual dataset directly. Lovenia et al. (2022) established basic multilingual code-switching ASR performance baselines from three differently pre-trained models: the XLSR multilingual pre-trained model (Conneau et al., 2020), and models fine-tuned for ASR using the English corpus and the Mandarin corpus of Common voice. They found the best performance in the Chinese pre-trained wav2vec 2.0 model. The model pre-fine-tuned on Mandarin Common Voice achieved the best performance out of the three. Lovenia et al. (2022) hypothesized that the slight performance discrepancy could be attributed to the fact that the ASCEND corpus consists of more than 50% of Mandarin speech. The larger vocabulary size in the model pre-trained in Mandarin provided it with performance advantages.

3.3 Improving ASR performance for wav2vec 2.0 models

While older ASR systems relied on separate acoustic models and language models to function, the newer trend in both the industry and academia is to leverage the much simpler structure of End-to-End systems. Wav2vec 2.0 has been the ASR model of choice for researchers in the field. We have seen impressive monolingual ASR performance using wav2vec 2.0 models. While wav2vec 2.0 model can decode the audio signal directly using CTC outputs with great performance, it is also possible to use a language model to rescore the outputs. Before using datasets to fine-tune the models, I want to test if language model decoding can help further reduce the WER for the Frisian wav2vec 2.0 model. Then, since Lovenia et al. (2022) showed that fine-tuning wav2vec 2.0 models with code-switching speech directly can improve the code-switching ASR performance, testing if such findings also hold for minority language pairs is the logical next step. In the meantime, I will also experiment if further fine-tuning the models with different data can also improve the performance.

3.3.1 Language Model Decoding

An informal search revealed that language models can help solve problems in producing more complete words. I used the transcription data from the Common Voice and FAME! corpora to train a language model with the KenLM library. Instead of making an interpolated language model like Bentum (2022), following their observation that the corpus text could be further cleaned, I decided that assembling a language model with a cleaned joint corpus text may yield better performance. The trained language model is tested in conjunction with the fine-tuned wav2vec 2.0 models from the XLSR model. We can see if the language model improves the performance for ASR tasks by comparing the testing results between the models with or without language models.

3.3.2 Fine-tuning with FAME! corpus

Similar to the models investigated in Lovenia et al. (2022), there are three different pre-trained models available for fine-tuning using the FAME! corpus, a multilingual corpus. The first one is the pre-trained wav2vec2-XLSR-53 model. The second one is a wav2vec 2.0 model fine-tuned for Frisian ASR tasks by Wietse de Vries[1]. The last one is a wav2vec 2.0 model fine-tuned for Dutch ASR tasks by Jonatas Grosman[2]. The latter two models are fine-tuned for monolingual ASR from the wav2vec 2.0 XLSR model using the Mozilla Common Voice corpus. To recap, the Mozilla Common Voice Corpus is a multilingual speech dataset that is produced and verified by native speakers of the respective language. The Frisian dataset consists of 1200 hours of recorded speech, 49 of which are verified. The Dutch dataset has 1200 hours of recorded speech, 99 of which are verified. To fine-tune the three different models, I initially used a Jupyter Notebook template for fine-tuning wav2vec 2.0 models for under-resourced language ASR tasks. But soon it became clear that using a Python script is much more efficient. Hence, I adapted the ASCEND training script to fine-tune the Frisian pre-trained wav2vec 2.0 model and the Dutch pre-trained wav2vec 2.0 model with the FAME! corpus. Because of the FAME! corpus was optimized for training in Kaldi, I conducted preliminary processing to load FAME! into a HuggingFace compatible dataset format.

After all fine-tuning steps were finished, I compare the performance metrics from the two different models. With the observations from the ASCEND training process, I hypothesized that, because FAME! contains more Frisian speech than Dutch speech, the performance of the Frisian fine-tuned model will be better in comparison. However, it should be noted that Mandarin and English output tokens are very different, Chinese cfileharacters versus the English alphabet. It is also possible that the performance advantages of a model that is more proficient in one language do not carry to the Frisian-Dutch speech.

Training details:

I followed Lovenia et al. (2022) and re-used the codebase provided by the paper. Adam optimizer and Connectionist Temporal Classification(CTC) loss were used. The training hyperparameters were inherited from the parent models. The models were fine-tuned using the GPU nodes with a single V100 GPU accelerator on the Peregrine computing cluster at the University of Groningen. The models were all trained for up to 100 epochs with early stopping.

[1] The pre-trained model could be found at https://huggingface.co/wietsedv/wav2vec2-large-xlsr-53-frisian

[2] The pre-trained model could be found at https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-dutch

3.4 Code-switching/loanword Classification

After finalizing the experiment on fine-tuning wav2vec 2.0 models on more data, the next step is to implement a language identification model to conduct the code-switching/loanword classification task. To do so, I extract the intermediate layer representations of the audio from the Common Voice dataset and train a simple model for language identification. While the majority of the data is only bilingual, Frisian, and Dutch, code-switching to other languages is also present in the datasets investigated. In the case of Frisian-Dutch code-switching, English terms could appear in the Frisian-Dutch speech. This issue, however, will not be accounted for during the LID model training because of limitations in the dataset.

Since the language identification system focuses on code-switching/loanword classification, it needs to operate on the word level instead of the higher utterance level. Hence I need to align the audio file with the transcriptions first to extract the timestamps for the word boundary for each word in the audio. Forced alignment is a method used to align transcriptions to audio files. The wav2vec2 model can also be used to conduct forced alignment thanks to its CTC architecture. Generating the start and end times of each letter in the utterance, I can extract the timestamps of the word boundaries and feed the corresponding audio segment into the LID model.

The language identification model is implemented in PyTorch with a single Long Short-Term Memory (LSTM) layer. The input features were generated using the wav2vec2-XLSR model. The LID model is trained using a combined Common Voice Dutch and Frisian dataset with a language label (locale as defined in the Common Voice format) as the output. The feature extraction module simply extracts an intermediate layer of the pre-trained ASR model for Frisian and Dutch. According to Pasad et al. (2021), layer 10 is chosen as the intermediate layer due to it being the layer that encodes the highest amount of phone identity. I encourage future researchers to test using LID task performance using different intermediate layers of the wav2vec 2.0 XLSR model. At the same time, a more sophisticated model architecture may achieve better performance, but due to time and resource constraints, I leave that to future researchers to investigate further.

3.4.1 Challenges

Language contact: Frisian and Dutch share a long history together (Markey, 1980; Steurs et al., 2022). And because of the shared history, Frisian and Dutch went through extensive language contact, and the low-level phonetic features of the two languages could appear very similar. Performing code-switch/loanword classification or even traditional language identification tasks for Frisian and Dutch may be significantly more difficult compared to more distant language pairs such as

Mandarin-English where the orthography and pronunciation are both significantly different.

Problems with the FCMC corpus: While the FCMC corpus offers a significant amount of new data for training, the corpus itself is not optimal for training ASR or TTS systems directly. Due to the difficulty in replicating the findings using the codebase provided by Bentum (2022), I decided to refrain from using the FCMC corpus for ASR system training. Future researchers may be able to get complete instructions from the programmers in the project to clean and pre-process the audio and transcription data so it would be better suited for spoken language applications.

3.5 Ethical Considerations

This thesis project utilizes multiple corpora in the speech recognition, code-switching, and multilingualism domain. The ASCEND Mandarin-English corpus (Lovenia et al., 2022) and the Common Voice Multilingual Speech Corpus were available as open-access datasets on HuggingFace. The SEAME Mandarin-English corpus (D.-C. Lyu et al., 2010) was purchased by a joint grant from the University of Groningen Library and Campus Fryslân. The FAME! Frisian-Dutch corpus (Yılmaz et al., 2016) and the FCMC Frisian-Dutch corpus (Bentum, 2022) were acquired specifically for research done in this thesis from the respective data controller of the corpora. This project conforms to the applicable licensing agreement for all the licensed datasets. Data processing was all completed on the Peregrine High-performance Computing cluster at the University of Groningen.

Code-switching is sometimes regarded as having less prestige. The study aims to not reinforce any stereotype associated with code-switching such as the language users doing so sound less intelligent or code-switching breaches language integrity. Considering the status of Frisian as a minority and under-resourced language, the focus on explaining code-switching is crucial. People code-switch for a variety of reasons, none of which is unnatural. To minimize this risk, chapter one provides a clear scientific background to the code-switching phenomenon. And the interpretation of the results from the current study aligns with the theoretical advancements in multilingualism research field.

Summary

The first section of this chapter examined the datasets used in the thesis in several different aspects from availability, content, and quality. The second section covered the process necessary for replicating the previous findings. Section three introduced methods for investigating the first research question in improving code-switching speech recognition performance for lower-resourced language pairs by fine-tuning end-to-end models directly with code-switching speech data. Section four outlined the steps for creating a language identification module based on the hidden embeddings generated from state-of-the-art end-to-end models. Finally, section five touched on ethical considerations in undertaking this thesis.

Chapter 4

Results & Discussion

This chapter presents the results from replicating experiments in previous literature and the findings from tthe original experiments outlined in the last chapter while also provides discussions on the implications of the results. The chapter is structured as follows: section one briefly shows the replication results. The second section presents findings on experimentations on the wav2vec 2.0 models with language model decoding and fine-tuning the models with code-switching dataset. Lastly, section three shows the experiment results of the language identification module training.

4.1 Replication

FAME! with Kaldi: I was able to match the performance of the Kaldi systems implemented by Bentum (2022) and Yılmaz et al. (2016) using the same recipes. However, the codebases used by the researchers for these projects are not very structured. It was not immediately clear how researchers referencing these literatures can easily replicate their results. One of the other interesting issues I ran into during replication was that the Kaldi framework has certain characteristics that affect testing the implementation on Peregrine, a High-performance Computing (HPC) cluster available at the University of Groningen. I was able to overcome the problems with certain drawbacks such as increased training time by leveraging the powerful discrete GPUs and tuning the training scripts according to the recommendations by the HPC staff.

ASCEND with wav2vec 2.0: After following the documentation provided by Lovenia et al. (2022), I was able to match the performance metrics presented in the paper. The codebase provided by the research team is easy to follow and well structured. Table 4.1 compares the replication results with the original experimental results. [1]

Replication experiment results				
Pre-training	**Validation**		**Test**	
language	**MER(%)**	**CER(%)**	**MER(%)**	**CER(%)**
Chinese	**30.61**	**25.84**	**25.66**	**21.93**
English	35.20	27.55	29.19	22.94
ASCEND original experiment results				
Pre-training	**Validation**		**Test**	
language	**MER(%)**	**CER(%)**	**MER(%)**	**CER(%)**
Chinese	**30.37**	**25.72**	**27.05**	**22.69**
English	35.77	28.07	28.72	22.78

Table 4.1: Result comparison between replication and the original paper.

While there are publicly available datasets and models for Frisian-Dutch bilingual ASR systems, it is not trivial for a new researcher to replicate their findings due to limited documentation. Some papers did provide a codebase hosted on GitHub. But the codebase is rather unorganized and does not include clear instructions on how a reader could easily reproduce the experiment. Even worse, several other conference proceedings on ASR implementations cited in this thesis did not provide any documentation or codebase for replication. For junior researchers in the field, it takes us too much effort just to find relevant resources to replicate existing studies. These kinds of practices hinder the healthy development of the research field. Replicability is crucial for any research project, especially for fields in computational technologies. Replicating the previous findings provides a concrete baseline performance of the ASR system before testing the new implementations.

4.2 Experiments on wav2vec 2.0

4.2.1 Language Model Decoding

Applying language model decoding in conjunction with the wav2vec 2.0 CTC output did not provide us with better performance. Instead, language model decoding destroyed the normal performance achieved by the model only fine-tuned on Common Voice Frisian data as shown in table 4.2.

The unexpected poor result is most likely caused by the extremely limited text data provided by the Common Voice dataset and the FAME! corpus. With only around 9000 utterances in Common Voice Frisian and similar amount of hours of

[1] The reproduced model can be found at https://huggingface.co/techsword/ASCEND-wav2vec2-chinese-zh-cn.

Model	WER (%)	CER (%)
Frisian wav2vec2 - No LM	15.91	4.17
Frisian wav2vec2 - With LM	100.86	83.46

Table 4.2: Comparison between direct decoding and language model decoding performance using the monolingual wav2vec 2.0 Frisian model.

utterances in FAME!, there is simply not enough text data for a successful language model. It is possible that training a language model on a combined dataset joining FAME!, FCMC, Common Voice along with other text data sources could provide real improvement. However, due to the time and resource constraints for this thesis project, I leave that to future researchers to investigate further.

4.2.2 FAME! with wav2vec

Encouraged by the performance from the ASCEND baseline ASR systems, I adopted the same methodology used by Lovenia et al. (2022) to fine-tune the wav2vec 2.0 models (the Frisian model, Dutch model, and the XLSR model) using the FAME! corpus. However, I found several challenges that came with the corpus. First, due to the multilingual nature, the corpus had many language tags surrounding code-switched words. The corpus also occasionally includes code-switching outside of Frisian and Dutch, e.g. switch to English. A simple text pre-processing script was used to remove the extra information from the transcriptions. Second, the size of the FAME! corpus is roughly on par with the size of the ASCEND corpus in terms of hours of speech, the amount of Frisian speech and Dutch speech is not quite balanced as mentioned in Chapter 3.

Initially, the logs produced by the training script showed some peculiar behaviors of the neural network during fine-tuning the XLSR directly with the FAME! corpus. The WER and validation loss from each training small batch decreased at first, showing a promising trend just like the ASCEND baseline. But, after only 10 epochs, both the WER and validation loss started ramping back up and finally stabilized at very high numbers. This could be the result of training hyperparameters not being optimized.I observed a similar trend in the training loss during the fine-tuning of the Frisian model, but the evaluation for the checkpoint around 7000 steps produced good WER results. Fine-tuning the Dutch ASR models did not produce similar results, instead, the loss and error rate followed a nice downward curve like the results in Lovenia et al. (2022).[2] ultimately the Frisian and Dutch models successfully completed training and produced sensible evaluation metrics otherwise, I believe the issue with the XLSR model was because the hyperparameters inherited

[2] The loss graphs can be found in the appendix.

from the XLSR were not optimal for fine-tuning with the present dataset. Due to various constraints, I did not test fine-tuning the XLSR model with different hyperparameters. But since fine-tuning XLSR directly with code-switching data produced worse performance compared to the models already fine-tuned for ASR tasks as shown in Lovenia et al. (2022), we can assume the same also holds for other language pairs. Even if the XLSR model was fine-tuned successfully, the ASR performance of the model would not be on par with the other two.

In sharp contrast to the language model decoding results, fine-tuning the Frisian and Dutch wav2vec 2.0 models with FAME! directly yielded great results.[3] As shown in table 4.3, the ASR model trained on FAME! data (wav2vec2-Frisian-FAME!) achieved significantly better performance compared to the other two monolingual models on the FAME! dataset. While its Dutch ASR performance is worse compared to the Dutch monolingual model, the Frisian monolingual performance only decreased slightly compared to the Frisian monolingual model. The WER metrics are comparable with the Mandarin-English code-switching ASR model trained in Lovenia et al. (2022). The encouraging result confirms the fist hypothesis and answers the first research question. With more training data and more sophisticated end-to-end models, I believe that the performance of code-switching ASR could improve significantly even without the aid of a LID module.

	wav2vec2-Frisian	**wav2vec2-Dutch**	**wav2vec2-Frisian-FAME!**
CV-Frisian	15.91	N/A	20.38
CV-Dutch	N/A	16.71	38.61
FAME!-CS	50.13	75.45	30.82

Table 4.3: WER performance comparison between the three fine-tuned wav2vec 2.0 models on three different datasets.

While I scrapped initial plans to use the FCMC corpus to fine-tune wav2vec 2.0 models due to the noisy data and the amount of pre-processing needed, using the corpus to pre-train the wav2vec 2.0 model instead of using it to fine-tune for downstream ASR tasks could potentially benefit the code-switching ASR performance of Frisian-Dutch speech. On the other hand, the same methodology used in this thesis and Lovenia et al. (2022) could still benefit from larger dataset size. If future researchers with more time and resources can better prepare the FCMC corpus for ASR training, the same methodology could bring on even better code-switching ASR performance.

[3] The fine-tuned models can be found respectively at https://huggingface.co/techsword/wav2vec-fame-frisian and https://huggingface.co/techsword/wav2vec-fame-dutch

4.3 Language Identification

To speed up the feature extraction and model training, only 5000 utterances from each language were used to train the language identification module. Trained on Common Voice Frisian and Dutch datasets, the model achieved only 49% accuracy after 80 epochs. The training only took 80 epochs because the training loss stabilized at a constant after 60 epochs after an upward trend. With no indications of training loss reducing, it does not make sense for me to continue spending more computational resources on training the model. Due to the poor performance in plain non-code-switching language identifications, testing the performance of the LID module on a word level using the FAME! corpus does not seem necessary at this point. If the LID model is not able to properly identify longer utterances such as full sentences in the Common Voice dataset, it is unlikely that it would get any reasonable results trying to conduct LID on the word level. This result suggests that the hypothesis for the second research question in this thesis is flawed. There are several possible explanations for the poor performance. Firstly, there may be simply not that many phonetic differences between Frisian and Dutch speech for the model to recognize that these are two different languages. Tseng et al. (2022) was able to get great LID performance for Mandarin and English because of the huge differences in the speech features of Mandarin and English. An alternative explanation is that layer 10 of the XLSR pre-trained model does not extract enough phonetic information for the LID task, or it encodes features that are illsuited for the LID task. At the same time, the training data size may have played a role as well: the size of the training data (5000 utterances per language in total) used to train this LID model is significantly smaller than the SEAME corpus used by Tseng et al. (2022).

4.4 Summary

This chapter discussed the results and the implications thereof from replicating previous findings and conducting original experiments. Some takeaway points are summarized here. Some datasets used in this thesis were not specifically made for ASR training purposes. Apart from the ASCEND corpus and the Common Voice corpus, the other datasets needed significant restructuring to be used for wav2vec2 training. During the restructuring process, and as also mentioned in Bentum (2022), it was clear that some of these datasets needed significant preprocessing and cleaning. Due to the time-consuming process in pre-processing FAME! and FCMC for training the LID model, I only chose to use the limited data of 5000 utterances per each language training set adapted from Common Voice. Given more time and resources it may be possible to extend this training set further. The same situation holds for further fine-tuning the end-to-end ASR model, too.

In light of the results presented, several discussion topics touched on different perspectives of the research project. As defined in the earlier chapters, a true multilingual ASR system should be able to handle code-switched utterances in addition to achieving great performance for monolingual utterances. The problem with the current deep learning techniques relies on large datasets to achieve state-of-the-art performance. Especially in the case of lower-resourced languages, for Frisian and Dutch, the language identification model suffers from a lack of data. However, reasonable code-switching ASR performance is achievable by fine-tuning a great performing monolingual wav2vec 2.0 model with a limited code-switching dataset. With the continuous development of end-to-end speech recognition models, it may not even be necessary for the future to use a LID module in conjunction with the speech recognition framework to achieve great ASR performance as shown in Lovenia et al. (2022) and the results presented in this chapter. The LID approach may lose its performance advantages for more language pairs soon. However, from a theoretical linguistics standpoint, having a LID model to test and confirm theories in sociolinguistics and multilingualism research may continue to be an interesting topic. Hence it may be worthwhile for future researchers to develop and test more robust LID approaches using different feature embeddings from different model architectures.

Chapter 5

Conclusion

This project explored using several techniques in improving code-switching speech recognition performance proven successful in dissimilar language pairs: language model decoding, and fine-tuning end-to-end speech recognition models with code-switching corpora for lower-resourced languages. At the same time, the thesis also investigated the possibility of using hidden layer representations from end-to-end ASR model as feature input to train a language identification model for a pair of very similar languages. The language model decoding approach proved ineffective due to the small size of text data available. The code-switching corpora fine-tuning approach achieved similar levels of performance as seen in dissimilar and high-resourced language pairs. The experiment on the language identification model revealed significant room for improvements for future researchers.

As language model decoding for wav2vec 2.0 models is generally successful for other higher-resourced languages, it is reasonable for us to believe that Frisian ASR with language model decoding can also improve given more text data. The wav2vec 2.0 fine-tuning technique using code-switching corpora is successful for both higher-resourced and lower-resourced language pairs. The success shows a promising future direction for developing true multilingual and code-switching capable ASR. Looking ahead, with more researchers laying their eyes on improving ASR system usability for a wider audience (multilingual speakers), more specialized code-switching datasets may become available for computational linguistic research. There is great prospect in further fine-tuning newer and more sophisticated end-to-end models with fresh datasets for multilingual and code-switching capable ASR. At the same time, while we have seen successes brought on by LID modules in improving multilingual and code-switching ASR performance for drastically different language pairs such as Mandarin and English, it may not be the ideal approach for closely related languages such as Frisian and Dutch. This is not to say that the experiment result in this thesis is discouraging. Rather, it opened up new pathways and brought new research topics to light. The difficulty in language identification for the two languages warrants significant future research in both computational linguistics and phonetics, for similar and dissimilar language pairs alike, in finding out the features that help human listeners differentiate between different dialects and languages. Testing alternative hidden layer outputs in end-to-end ASR neural networks for different tasks may also help fuel future developments in this area. Where words come from may not be relevant for building code-switching automatic speech recognition systems, but it does matter for other linguistic research.

Appendix

Training visualizations

Attached below are training visulization graphs for fine-tuning wav2vec 2.0 models.

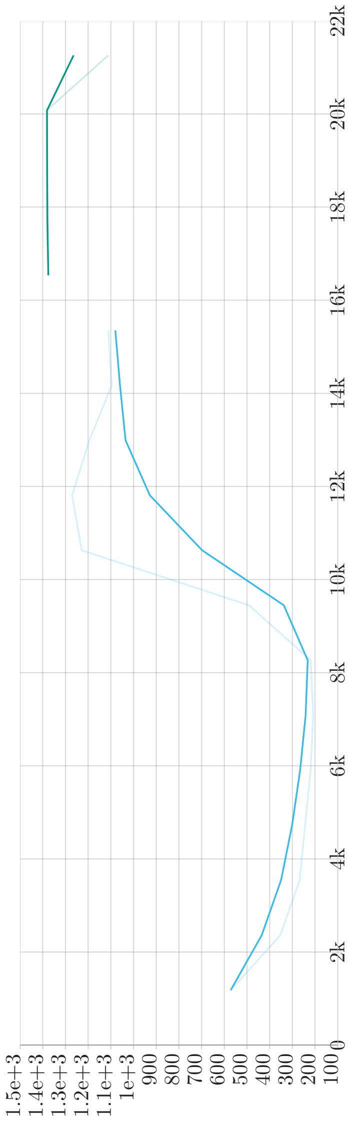

Figure 1: wav2vec2-large-xlsr-53 Eval Loss Graph

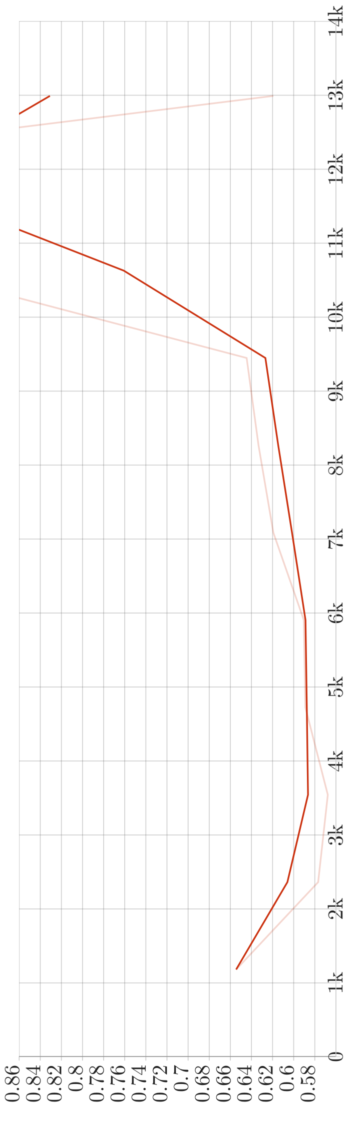

Figure 2: wav2vec2-large-xlsr-53-frisian Eval Loss Graph

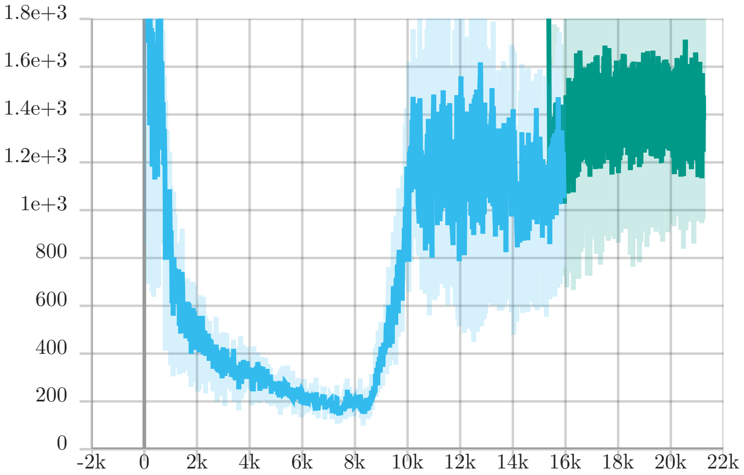

Figure 3: wav2vec2-large-xlsr-53 Train Loss Graph

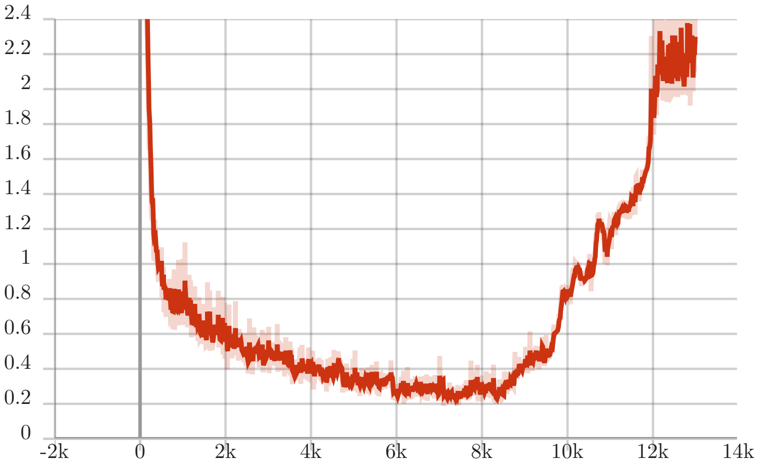

Figure 4: wav2vec2-large-xlsr-53-frisian Train Loss Graph

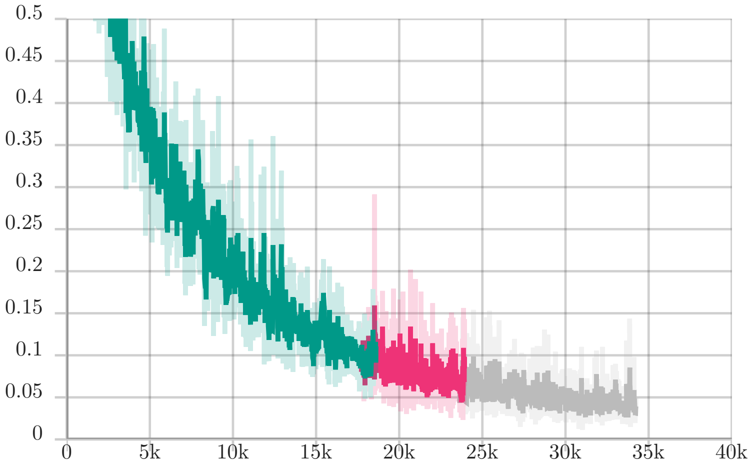

Figure 5: wav2vec2-large-xlsr-53-dutch Train Loss Graph

Bibliography

Adel, H., Vu, N. T., Kraus, F., Schlippe, T., Li, H., & Schultz, T. (2013). Recurrent neural network language modeling for code switching conversational speech. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 8411–8415. https://doi.org/10.1109/ICASSP.2013.6639306

Ambikairajah, E., Li, H., Wang, L., Yin, B., & Sethu, V. (2011). Language Identification: A Tutorial. *IEEE Circuits and Systems Magazine*, *11*(2), 82–108. https://doi.org/10.1109/MCAS.2011.941081

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., LeGresley, P., Lin, L., ... Zhu, Z. (2015, December 8). *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.* Retrieved April 26, 2022, from http://arxiv.org/abs/1512.02595

Appel, R., & Muysken, P. (1987). *Language contact and bilingualism.* E. Arnold.

Appel, R., & Muysken, P. (2005). Code switching and code mixing. In *Language Contact and Bilingualism* (pp. 117–128). Amsterdam University Press.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020, March 5). *Common Voice: A Massively-Multilingual Speech Corpus.* Retrieved June 15, 2022, from http://arxiv.org/abs/1912.06670

Badiola, L., Delgado, R., Sande, A., & Stefanich, S. (2018). Code-switching attitudes and their effects on acceptability judgment tasks. *Linguistic Approaches to Bilingualism*, *8*(1), 5–24. https://doi.org/10.1075/lab.16006.bad

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020, October 22). *Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.* Retrieved March 22, 2022, from http://arxiv.org/abs/2006.11477

Bentum, M. (2022). A Speech Recognizer for Frisian/Dutch Council Meetings, 7.

Bosma, E., & Blom, E. (2019). A code-switching asymmetry in bilingual children: Code-switching from Dutch to Frisian requires more cognitive control than code-switching from Frisian to Dutch. *International Journal of Bilingualism*, *23*(6), 1431–1447. https://doi.org/10.1177/1367006918798972

Bullock, B. E. (2009). Phonetic reflexes of code-switching. In A. J. Toribio & B. E. Bullock (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp. 163–181). Cambridge University Press. https://doi.org/10.1017/CBO9780511576331.011

Chang, C.-T., Chuang, S.-P., & Lee, H.-Y. (2019, June 19). *Code-switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation.* Retrieved April 14, 2022, from http://arxiv.org/abs/1811.02356

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020, December 15). *Unsupervised Cross-lingual Representation Learning for Speech Recognition* (arXiv:2006.13979). https://doi.org/10.48550/arXiv.2006.13979

de Vries, W., Bartelds, M., Nissim, M., & Wieling, M. (2021). Adapting Monolingual Models: Data can be Scarce when Language Similarity is High. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 4901–4907. https://doi.org/10.18653/v1/2021.findings-acl.433

Donaldson, B. (1983). *Dutch. A linguistic history of Holland and Belgium.*

Dong, L., Xu, S., & Xu, B. (2018). Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5884–5888. https://doi.org/10.1109/ICASSP.2018.8462506

Dorleijn, M., & Nortier, J. (2009). Code-switching and the internet. In B. E. Bullock & A. J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp. 127–141). Cambridge University Press. https://doi.org/10.1017/CBO9780511576331.009

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2022). *Ethnologue: Languages of the World* (Twenty-fifth). SIL International. http://www.ethnologue.com.

Elias, V., McKinnon, S., & Milla-Muñoz, Á. (2017). The Effects of Code-Switching and Lexical Stress on Vowel Quality and Duration of Heritage Speakers of Spanish. *Languages*, *2*(4), 29. https://doi.org/10.3390/languages2040029

Fan, Z., Li, M., Zhou, S., & Xu, B. (2021). Exploring wav2vec 2.0 on Speaker Verification and Language Identification. *Interspeech 2021*, 1509–1513. https://doi.org/10.21437/Interspeech.2021-1280

Gooskens, C., & Heeringa, W. (2012). The Position of Frisian in the Germanic Language Area.

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81. https://doi.org/10.1017/S1366728998000133

Hannun, A. (2021, July 30). The History of Speech Recognition to the Year 2030. Retrieved July 28, 2022, from http://arxiv.org/abs/2108.00084

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014, December 19). *Deep Speech: Scaling up end-to-end speech recognition* (arXiv:1412.5567). https://doi.org/10.48550/arXiv.1412.5567

Huang, Z., Wang, P., Wang, J., Miao, H., Xu, J., & Zhang, P. (2021). Improving Transformer Based End-to-End Code-Switching Speech Recognition Using Language Identification. *Applied Sciences*, *11*(19), 9106. https://doi.org/10.3390/app11199106

Language plan Frisian 2030 –Frisian, for now and later. (n.d.). Retrieved July 28, 2022, from https://taalplan.frl/

Li, K., Li, J., Ye, G., Zhao, R., & Gong, Y. (2019). Towards Code-switching ASR for End-to-end CTC Models. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6076–6080. https://doi.org/10.1109/ICASSP.2019.8683223

Liu, D., Xu, J., Zhang, P., & Yan, Y. (2021). A unified system for multilingual speech recognition and language identification. *Speech Communication*, *127*, 17–28. https://doi.org/10.1016/j.specom.2020.12.008

Liu, H., Perera, L. P. G., Zhang, X., Dauwels, J., Khong, A. W., Khudanpur, S., & Styles, S. J. (2021). End-to-End Language Diarization for Bilingual Code-Switching Speech. *Interspeech 2021*, 1489–1493. https://doi.org/10.21437/Interspeech.2021-82

Liu, H. (2019). Attitudes Toward Different Types of Chinese-English Code-Switching. *SAGE Open*, *9*(2). https://doi.org/10.1177/2158244019853920

Lovenia, H., Cahyawijaya, S., Winata, G. I., Xu, P., Yan, X., Liu, Z., Frieske, R., Yu, T., Dai, W., Barezi, E. J., Chen, Q., Ma, X., Shi, B. E., & Fung, P. (2022, April 28). *ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation*. Retrieved May 2, 2022, from http://arxiv.org/abs/2112.06223

Lu, J.-Y. (1991). Code-switching between Mandarin and English. *World Englishes*, *10*(2), 139–151. https://doi.org/10.1111/j.1467-971X.1991.tb00147.x

Lyu, D., & Lyu, R.-Y. (2008). Language identification on code-switching utterances using multiple cues. *undefined*. Retrieved March 10, 2022, from https://www.semanticscholar.org/paper/An-integrated-language-identification-for-code-and-Mabokela-Manamela/9d0a4c5f158c9222f7d9918eb43703347f5a6a0c

Lyu, D.-C., Tan, T.-P., Chng, E. S., & Li, H. (2010). SEAME: A Mandarin-English code-switching speech corpus in south-east asia. *Interspeech 2010*, 1986–1989. https://doi.org/10.21437/Interspeech.2010-563

Lyu, D.-C., Tan, T.-P., Chng, E.-S., & Li, H. (2015). Mandarin—English code-switching speech corpus in South-East Asia: SEAME. *Language Resources and Evaluation*, *49*(3), 581–600.

Mager, M., Çetinoğlu, Ö., & Kann, K. (2019, April 3). *Subword-Level Language Identification for Intra-Word Code-Switching*. Retrieved March 10, 2022, from http://arxiv.org/abs/1904.01989

Markey, T. L. (1980). *Frisian*. De Gruyter Mouton. https://doi.org/10.1515/9783110815719

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Muldner, K., Hoiting, L., Sanger, L., Blumenfeld, L., & Toivonen, I. (2019). The phonetics of code-switched vowels. *International Journal of Bilingualism*, *23*(1), 37–52. https://doi.org/10.1177/1367006917709093

Muysken, P. (1995). Code-switching and grammatical theory. In L. Milroy & P. Muysken (Eds.), *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching* (pp. 177–198). Cambridge University Press. https://doi.org/10.1017/CBO9780511620867.009

Myers-Scotton, C. (1993). *Duelling languages: Grammatical structure in codeswitching*. Clarendon Press ; Oxford University Press.

Myers-Scotton, C. (2002a). *Contact linguistics: Bilingual encounters and grammatical outcomes*. Oxford University Press.

Myers-Scotton, C. (2002b). Frequency and intentionality in (un)marked choices in codeswitching: ”this is a 24-hour country”. *International Journal of Bilin-*

gualism, *6*(2), 205–. Retrieved April 27, 2020, from http://link.gale.com/apps/doc/A91039986/AONE?u=s8888903&sid=zotero&xid=8aa9a9e0

Myers-Scotton, C., & Jake, J. (2009). A universal model of code-switching and bilingual language processing and production. In B. E. Bullock & A. J. Toribio (Eds.), *The Cambridge Handbook of Linguistic Code-switching* (pp. 336–357). Cambridge University Press. https://doi.org/10.1017/CBO9780511576331.020

Olson, D. J. (2016). The role of code-switching and language context in bilingual phonetic transfer. *Journal of the International Phonetic Association*, *46*(3), 263–285. https://doi.org/10.1017/S0025100315000468

O' Shaughnessy, D. (2008). Automatic speech recognition: History, methods and challenges. *Pattern Recognition*, *41*(10), 2965–2979. https://doi.org/10.1016/j.patcog.2008.05.008

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., & Auli, M. (2019). Fairseq: A Fast, Extensible Toolkit for Sequence Modeling, 48–53. https://doi.org/10.18653/v1/N19-4009

Parafita Couto, M. C., & Gullberg, M. (2019). Code-switching within the noun phrase: Evidence from three corpora. *International Journal of Bilingualism*, *23*(2), 695–714. https://doi.org/10.1177/1367006917729543

Pasad, A., Chou, J.-C., & Livescu, K. (2021, October 8). Layer-wise Analysis of a Self-supervised Speech Representation Model. Retrieved July 20, 2022, from http://arxiv.org/abs/2107.04734

Piccinini, P., & Arvaniti, A. (2015). Voice Onset Time in Spanish-English Spontaneous Code-Switching. *Journal of Phonetics*, *52*, 121–137. https://doi.org/10.1016/j.wocn.2015.07.004

Piccinini, P., & Garellek, M. (2014). Prosodic Cues to Monolingual versus Code-switching Sentences in English and Spanish. *Speech Prosody 2014*, 885–889. https://doi.org/10.21437/SpeechProsody.2014-166

Poarch, G. J., & Bialystok, E. (2015). Bilingualism as a model for multitasking. *Developmental Review*, *35*, 113–124. https://doi.org/10.1016/j.dr.2014.12.003

Poplack, S. (1980). Sometimes I' ll start a sentence in Spanish Y TERMINO EN ESPAÑOL: Toward a typology of code-switching 1. *Linguistics*, *18*, 581–618. https://doi.org/10.1515/ling.1980.18.7-8.581

Poplack, S., Sankoff, D., & Miller, C. (1988). The social correlates and linguistic processes of lexical borrowing and assimilation. *26*(1), 47–104. https://doi.org/10.1515/ling.1988.26.1.47

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıˇcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit, 4.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). MLS: A Large-Scale Multilingual Dataset for Speech Research. *Interspeech 2020*, 2757–2761. https://doi.org/10.21437/Interspeech.2020-2826

Qian, Y.-m., & Xiang, X. (2019). Binary neural networks for speech recognition. *Frontiers of Information Technology & Electronic Engineering*, *20*(5), 701–715. https://doi.org/10.1631/FITEE.1800469

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2015, July 31). *FAVE: Speaker fix.* https://doi.org/10.5281/zenodo.22281

Shen, A., Gahl, S., & Johnson, K. (2020). Didn't hear that coming: Effects of withholding phonetic cues to code-switching. *Bilingualism: Language and Cognition*, *23*(5), 1020–1031. https://doi.org/10.1017/S1366728919000877

Shen, H.-P., Wu, C.-H., Yang, Y.-T., & Hsu, C. (2011). CECOS: A Chinese-English code-switching speech database. *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*. https://doi.org/10.1109/ICSDA.2011.6085992

Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., & Khudanpur, S. (2018). Spoken Language Recognition using X-vectors. *The Speaker and Language Recognition Workshop (Odyssey 2018)*, 105–111. https://doi.org/10.21437/Odyssey.2018-15

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Steurs, F., Vandeghinste, V., & Daelemans, W. (2022). Report on the Dutch Language, 23.

Tseng, L.-H., Fu, Y.-K., Chang, H.-J., & Lee, H.-y. (2022). Mandarin-English Code-switching Speech Recognition with Self-supervised Speech Representation Models, 6.

Voigt, R., Jurafsky, D., & Sumner, M. (2016). Between- and Within-Speaker Effects of Bilingualism on F0 Variation, 1122–1126. https://doi.org/10.21437/Interspeech.2016-1506

Wei, R., & Su, J. (2012). The statistics of English in China: An analysis of the best available data from government sources. *English Today*, *28*(3), 10–14. https://doi.org/10.1017/S0266078412000235

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yang, Y.-Y., Hira, M., Ni, Z., Chourdia, A., Astafurov, A., Chen, C., Yeh, C.-F., Puhrsch, C., Pollack, D., Genzel, D., Greenberg, D., Yang, E. Z., Lian, J., Mahadeokar, J., Hwang, J., Chen, J., Goldsborough, P., Roy, P., Narenthiran, S., … Shi, Y. (2022, February 16). *TorchAudio: Building Blocks for Audio and Speech Processing.* Retrieved May 8, 2022, from http://arxiv.org/abs/2110.15018

Yilmaz, E., Van Den Heuvel, H., & Van Leeuwen, D. (2018). Code-Switching Detection with Data-Augmented Acoustic and Language Models. *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, 127–131. https://doi.org/10.21437/SLTU.2018-27

Yılmaz, E., Cohen, S., Yue, X., van Leeuwen, D., & Li, H. (2019, June 28). *Multi-Graph Decoding for Code-Switching ASR*. Retrieved March 28, 2022, from http://arxiv.org/abs/1906.07523

Yılmaz, E., van den Heuvel, H., & van Leeuwen, D. (2016). Investigating Bilingual Deep Neural Networks for Automatic Recognition of Code-switching Frisian Speech. *Procedia Computer Science*, *81*, 159–166. https://doi.org/10.1016/j.procs.2016.04.044

Zhang, H. (2012). Code-switching Speech Detection Method by Combination of Language and Acoustic Information, 4.